# Mapping of oncology regimen data through an LLM-enhanced pipeline

**Tatsiana Skuhareuskaya[1], Mikita Salavei[1], Qi Yang[2], Maria Khitrun[1], Vlad Korsik[1]**

**[1]Odysseus, an EPAM Company; [2]Analyst, Data Strategy, Access, and Enablement (DSAE), IQVIA Inc.**

## Background

In observational data in general and in Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) specifically, data on oncology treatment regimens is a highly sought after asset [1]. Standardizing the transformation of such data, in both Extract-Transform-Load (ETL) and semantic mapping processes, is a challenging task which requires logic development in terms of both temporal logic and drug combinations. OHDSI Community is developing a set of tools that are helpful in Regimen derivation from OMOP Data Tables, with Artemis [2] as a prominent example. However, Source Data Assets may contain already aggregated and curated information about such complex therapeutic exposures. In this case, semantic mapping and direct landing to the Episode table seems an attractive option. Such a direct table population may help to benchmark the community-build abstraction tools, as well as serve as backbone for episode_event table population via reverse engineering. Basic lexical matches, for example produced by USAGI [3], are often too superficial and require substantial manual review effort. Here we describe an LLM-based approach for regimen mapping, from text preprocessing to mapping prioritization.

## Methods

We used OncEMR freetext-encoded regimen data as source, and HemOnc Standardized vocabulary, including synonyms of the concepts, as the target ontology. Although regimens with similar drugs prescribed in different dosing or timing patterns are clinically considered separate regimens, since the target regimen ontology used in OMOP CDM is HemOnc and the OMOP-adopted HemOnc version contains only drug-level regimen descriptions, we aimed at preprocessing the OncEMR descriptions to the level of drugs included. Through iterative testing, we arrived at a preprocessing pipeline which ensured consistency along with data preservation. We utilize embedding potential of 3 large language models: SapBERT [4], BioLORD-2023 [5] and PubMedBERT [6] to generate multidimensional vectors for preprocessed sources and targets. The processed OncEMR source data, along with the HemOnc vocabulary and concept synonyms—each represented via learned embeddings—were then assessed for semantic closeness metrics (Cosine similarity and Euclidean distance), with top-10 candidates chosen for subject matter expert (SME) review. SMEs were medical doctors experienced in working with medical ontologies, including HemOnc.

## Results

After evaluating several text preprocessing strategies, we identified an optimal pipeline that included tokenization, removal of custom stop words, and elimination of single-character tokens.

Initial experiments using Cosine similarity for entity matching produced overly dense associations, complicating mapping candidate choice and making it unreliable. Consequently, we selected Euclidean distance as the preferred similarity metric, which resulted in more robust and discriminative mappings.

This approach yielded mappings from 19,128 distinct codes to 960 unique HemOnc regimens, while effectively filtering out supportive care and irrelevant ("junk") codes, resulting in 2,735 codes with no corresponding regimen (zero matches).

The final output was grouped by source code, ranked by the number of LLMs that produced the mapping, and formatted for expert review in a human-readable manner. To assess mapping accuracy, we calculated the percentage of LLM-generated mappings that were retained without modification after expert validation. Mapping accuracy achieved 79.6% reliability for cases where the candidate target was suggested by all three embedding models. For only two models, the LLM success rate was 20.4%. We also compared the mappings suggested by the vectorization approach with those created by USAGI. For USAGI mappings, approximately 1 in 2 mappings had to be changed by the reviewer compared to 1 in 3 for the LLM-enhanced approach.

Out of the regimens mapped using this approach, 362 were used only in one corresponding nosology. The list of top-20 cancers with several corresponding mapped regimens is presented in Table 1.

**Table 1.** Top 20 cancer types with their regimens mapped using the LLM-enhanced approach

| concept_name of cancer | Count of mapped regimens |
|---|---|
| Breast cancer | 174 |
| Colorectal cancer | 88 |
| ERBB2 Breast cancer | 85 |
| Non-small cell lung cancer | 78 |
| Diffuse large B-cell lymphoma | 77 |
| Multiple myeloma | 71 |
| Non-small cell lung cancer nonsquamous | 48 |
| Ovarian cancer | 48 |
| Mantle cell lymphoma | 46 |

| | |
|---|---|
| Gastric cancer | 46 |
| Cervical cancer | 45 |
| Follicular lymphoma | 44 |
| Chronic lymphocytic leukemia | 43 |
| Melanoma | 35 |
| Small cell lung cancer | 34 |
| TNBC Breast cancer | 34 |
| Head and neck cancer | 32 |
| Urothelial carcinoma | 30 |
| Classical Hodgkin lymphoma | 30 |
| Hepatocellular carcinoma | 29 |

**Conclusion**

Our approach enables the efficient generation of large-scale relationships for custom concepts within the complex domain of treatment regimens. Future directions may include the integration of advanced matching algorithms and the development of complementary accuracy metrics, further enhancing the precision and utility of AI-assisted ontology engineering.

**References:**


1. Warner JL, Dymshyts D, Reich CG, et al. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. J Biomed Inform. 2019;96:103239. doi:10.1016/j.jbi.2019.103239
2. https://github.com/OHDSI/ARTEMIS
3. https://github.com/OHDSI/Usagi
4. https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext
5. https://huggingface.co/FremyCompany/BioLORD-2023
6. https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract